VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ
DEPARTMENT OF BIOMEDICAL ENGINEERING

KONZERVACE POZICE GENŮ V BAKTERIÁLNÍCH GENOMECH
GENE ORDER CONSERVATION IN BACTERIAL GENOMES

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **Tereza Martinková**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Denisa Maděránková, Ph.D.**
SUPERVISOR

BRNO 2018

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní obor **Biomedicínská technika a bioinformatika**
Ústav biomedicínského inženýrství

Studentka: Tereza Martinková ***ID:*** 183354
Ročník: 3 ***Akademický rok:*** 2017/18

NÁZEV TÉMATU:

Konzervace pozice genů v bakteriálních genomech

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši na téma konzervace pozice genů (synteny bloky) v prokaryotních genomech a na téma komparativní genomiky založené na porovnání pozice genů. 2) Sestavte rozsáhlý dataset anotovaných bakteriálních genomů, který bude sloužit k analýze jejich fylogenetických vztahů porovnáním vzájemných pozic genů. 3) V libovolném programovém prostředí vytvořte funkci pro výpočet pozičního profilu vybraných genů anotovaného genomu vůči vybranému referenčnímu genomu. 4) Navrhněte a implementujte metodu porovnání pozičního profilu vybraných genů s cílem vyhodnocení fylogenetických vztahů mezi bakteriálními genomy. 5) Proveďte analýzu sestaveného datasetu a výsledky diskutujte a porovnejte s referenčním fylogenetickým stromem.

DOPORUČENÁ LITERATURA:

[1] MAHADEVAN, P. a D. SETO. Rapid pair-wise synteny analysis of large bacterial genomes using web-based GeneOrder 4.0. BMC Research Notes. 2010, 3:41.
[2] SODERLUND, C., NELSON, W., SHOEMAKER, A. SyMAP: A system for discovering and viewing syntenic regions of FPC maps. Genome Research, 2006, roč. 16, s. 1159-1168.

Termín zadání: 5. 2. 2018 ***Termín odevzdání:*** 8. 8. 2018

Vedoucí práce: Ing. Denisa Maděránková, Ph.D.

prof. Ing. Ivo Provazník, Ph.D.
předseda oborové rady

Abstrakt

Teoretická část práce se zabývá základními pojmy, jako je bakteriální genom, komparativní genomika, a hlavně synteny bloky. Je zde vysvětleno, co synteny je a v čem spočívá její důležitost. Dále je v teoretické části zmíněn GenBank formát, jeho obsah a využití. Praktická část je zaměřena na vyhledávání podobností v sekvencích DNA referenční bakterie s vybranou bakterií, dále jejich setřídění pomocí hladového algoritmu a vyhodnocení výsledků.

Klíčová slova

Breakpointová metoda, CDS, DNA, fylogenetika, gen, genom, komparativní genomika, synteny

Abstract

Theoretical part of the thesis deals with basic concepts such as bacterial genome, comparative genomics and mainly synteny blocks. Here is explained what synteny is and what is its importance. In the theoretical part, the GenBank format is also mentioned, its content and usage. The practical part is focused on searching similarities in DNA sequences of reference bacteria with selected bacteria, heir sorting by means of hungry algorithm and evaluation of the results.

Keywords

Breakpoint method, CDS, DNA, phylogenetics, gene, genome, comparative genomics, synteny

Bibliografická citace:

MARTINKOVÁ, T. *Konzervace pozice genů v bakteriálních genomech*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2018. 36s. Vedoucí práce: Ing. Denisa Maděránková, Ph.D.

Prohlášení

„Prohlašuji, že svou bakalářskou práci na téma Konzervace pozice genů v bakteriálních genomech jsem vypracovala samostatně pod vedením vedoucí bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce. Jako autor uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestně právních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne **8. srpna 2018** ……

podpis autorky

Poděkování

Děkuji vedoucí bakalářské práce Ing. Denise Maděránkové, Ph.D. za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé bakalářské práce.

V Brně dne **8. srpna 2018** ……

podpis autorky

Obsah

Seznam obrázků

Seznam tabulek

ÚVOD

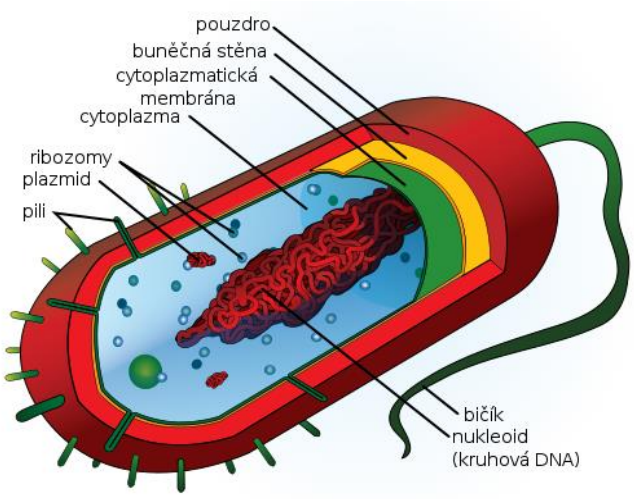
Bakterie jsou jednobuněčné prokaryotické organismy, jejich genomy jsou zpravidla mnohem menší než genomy eukaryotických buněk. Díky tomu jsou bakterie velice přínosné v genetice, slouží jako modelové organismy díky nejjednodušší manipulaci s jejich geny. Nejčastěji zkoumanými bakteriemi jsou patogenní, kde se hledají další možnosti léčby a obrany proti těmto bakteriím. Jedním z odvětví genomiky je komparativní genomika – obor zajímající se o podobnosti a vztahy mezi geny, chromozomy nebo celými genomy. Využívá se pro porovnání různých organismů, například díky komparativní genomice byla zjištěna podobnost lidského genomu s hlodavčím. Zkoumáním bakteriálního genomu získáváme hodně informací o evolučních procesech. O nich více prozradí i tzv. synteny bloky, což jsou vzájemně podobné oblasti v genomech rozdílných organismů, kde se porovnává pořadí synteny neboli skupinu homologních sekvencí, které se ve chvíli porovnání vyskytují zároveň.

V teoretické části práce je kromě synteny popsán i bakteriální genom, anotace genu nebo rozsáhlý formát genbank, který obsahuje velké množství informací o daném záznamu sekvence, která je v něm uložena. Teoretickou část uzavírá krátké shrnutí o použitých bakteriích, kde byla podobnost vybírána na základě jejich gramnegativního barvení.

Praktická část této práce je věnována přímému vyhledávání podobností mezi jednotlivými bakteriemi. Využívá shod v názvech genů a lokální zarovnání sekvence aminokyselin. Vyhledávání shod slouží k zjištění množství změn v pořadí genů mezi referenční a vybranou bakterií. Pořadí genů je proto setříděno jedním z hladových algoritmů a je zjištěn počet kroků, který byl potřebný k seřazení pozičního vektoru. Počet kroků metody určí evoluční vzdálenost mezi referenční a vybranou bakterií, s jeho pomocí se spočítá *p*-distance a vyhodnotí se úspěšnost metody.

1 BAKTERIE

Bakterie se řadí mezi prokaryotní organismy. Jsou tvořeny prokaryotní (chybí zde jádro) buňkou s jediným chromozomem kruhovitého tvaru umístěným volně v cytoplazmě. Postrádají i složitý systém membránových organel, např. neobsahují mitochondrie a vakuoly, ovšem jejich funkce je nahrazována. Metabolické děje u bakterií probíhají volně v cytoplazmě. Molekuly DNA jsou v prokaryotických buňkách uloženy buď v hlavním chromozomu nebo v plazmidech – malé molekuly DNA, též kružnicového tvaru (Obrázek 1.1). Neprobíhají zde meiotické ani mitotické děje, bakterie se rozmnožují nepohlavně příčným dělením. Za krátkou dobu je tedy možné vytvořit mnohamilionové populace, kde se zvyšuje šance na pozorování velmi vzácných mutací, které by se u menších populací téměř jistě neprojevily. Bakterie se v genetice využívají hlavně kvůli rychlosti rozmnožování. Jsou použity jako modelové organismy pro výzkum, jejich genomy patří mezi nejkratší, je s nimi tedy o něco snadnější manipulace. Cílenými mutacemi DNA se zjišťuje funkce genů, enzymů nebo metabolických cest. Nejčastěji se jako modelový organismus využívá *Escherichia coli*. [1][2]



Obrázek 1.1 Stavba prokaryotické buňky

1.1 Bakteriální genom

Genetický materiál bakterií se skládá z hlavního chromozomu a vedlejších elementů, plazmidů a epizomů. Chromozom je tvořen jedinou dvoušroubovicí stočenou do kruhu – superhelicita (nadšroubovicové vinutí). Je též nazýván jako nukleoid. Eukaryotické buňky mají své chromozomy zabaleny pomocí proteinů nazývaných histony, prokaryotické buňce histony chybí, ale vyskytují se zde jiné proteiny analogické

histonům. Chromozom je haploidní a pouze jeho malá část je nekódující, obsahuje jednoduché strukturní geny. [2][3][4]

Replikace bakteriálních chromozomů má tři hlavní fáze – zahájení, prodloužení a ukončení. Počátek replikace, anglicky origin of replication, zkráceně oriC, je konkrétní sekvence genomu, kde začíná fáze zahájení a replikace z něj probíhá souměrně a symetricky v opačných směrech. Struktura oriC se liší druh od druhu, ale některé vlastnosti jsou společné. Jedna ze společných charakteristik oriC je vysoký obsah AT nukleotidů. Na oriC je navázán předreplikační komplex, což je proteinový komplex tvořící se v průběhu zahájení replikace. U bakterií je hlavní složkou prereplikačního komplexu protein DnaA, který je replikační iniciační faktor a zahájení iniciační fáze je dáno jeho koncentrací. DnaA se váže na specifická místa v oriC, čtyři úseky o délce 9 bp se sekvencemi 5‘-TTAT(C nebo A)CA(C nebo A)A-3‘. Přibližně 10 molekul DnaA je vázáno na tyto úseky, po tomto navázání se DNA začíná rozvolňovat. Aktivní forma DnaA spotřebovává energii ATP. Hlavním enzymem replikace je DNA polymeráza III, syntéza hlavních řetězců začíná syntézou RNA primeru v oriC, celou reakci poté katalyzuje DNA primáza. Syntéza hlavního řetězce probíhá kontinuálně, ovšem na zaostávajícím řetězci vznikají tzv. Okazakiho fragmenty. Vznikají díky neschopnosti DNA polymerázy syntetizovat nové nukleotidy ve směru 3‘-5‘. Jsou to krátké, nově vytvořené úseky DNA, komplementární k zaostávajícímu řetězci a za pomoci DNA ligázy se spojí v jeden celek. [5][6][7]

Bakteriální transkripce je proces, ve kterém vzniká z bakteriální DNA mediátorová RNA, provádí se za pomoci RNA polymerázy. Obvykle se transkripce zahajuje navázáním RNA polymerázy na promotor, což je specifický úsek DNA. Na bakteriální promotor se naváže dočasně RNA polymeráza, spíše jedna z jejích podjednotek, tzv. sigma faktor. Sigma faktor umožňuje iniciaci transkripce a výběr promotoru, u bakterií je důležitým prvkem při regulaci genové exprese. Bakterie mají více sigma faktorů s rozdílnými funkcemi. Jeden je určen na provoz buňky, další umožňují bakteriím měnit transkripční program. Promotor obsahuje dvě vysoce konzervované sekvence, které sigma faktor rozeznává. Pro nejčastěji se vyskytující sigma faktor σ^{70} (RpoD), se jedná o tzv. Pribnow box na pozici -10. Pribnow box je sekvence 6 nukleotidů (TATAAT), nazýván také jako -10 sekvence, jelikož je umístěn zhruba 10 párů bazí před místem iniciování transkripce. Jeho funkce je podobná jako funkce TATA boxu u eukaryotických buněk. [3][4]

1.1.1 Plazmidy

Krom hlavního chromozomu je genetická informace ve většině bakterií uložena ještě v plazmidech a epizomech. Liší se od sebe velikostně (epizomy jsou vetší) a především tím, že plazmidy se nemohou spojit s chromozomální DNA. Plazmid nese všechny potřebné informace pro jeho nezávislou replikaci. I přes to, že plazmidy nejsou nezbytné pro přežití buňky, jejich přítomnost poskytuje jisté výhody, mnohé nesou geny, které poskytují rezistenci vůči antibiotikům nebo těžkým kovům. Tyto vlastnosti mohou být předány jiné bakterii. [8][9][10]

Plazmidy též hrají důležitou roli při konjugaci (spojení dvou bakterií za účelem výměny genetické informace), navíc některé výzkumy poukazují na skutečnost, že některé charakteristické vlastnosti bakterií využívající se v medicíně nebo průmyslu jsou právě díky genům, které nesou plazmidy. Jsou snadno izolovatelné, dají se znovu zavést do jiných bakterií, proto mají zásadní význam při studiu chromozomového přeskupení v bakteriích. [8][9]

V buňce se vyskytují různé druhy plazmidů s rozdílnými funkcemi. R-plazmidy (rezistenční) nesou geny, které zodpovídají za rezistenci buňky k antibiotikům a jiným antibakteriálním látkám. Některé R-plazmidy podmiňují schopnost konjugace bakterií. Tato vlastnost má velký význam při šíření genů rezistentních k antibiotikům v populacích patogenních bakterií. Ovšem v medicíně způsobily R-plazmidy veliký problém, kvůli nadměrnému používání antibiotik se mnohé bakterie staly proti nim rezistentní. [1][8][11]

1.1.2 Epizomy

Epizomem se rozumí přídavný genetický materiál u prokaryotních organismů. Mohou být uloženy volně nebo jako součást chromozomu. Epizomy (doslovně „additional body“, tedy „přídavné tělo“) jsou pro buňku postradatelné, ovšem jejich přítomnost jí může zajistit určité výhody, jako např. lepší adaptovatelnost a reakci na podmínky prostředí. Tyto genetické elementy mohou existovat ve dvou možných stavech, v prvním, integrovaném stavu, je epizom spojen s bodem nebo malou oblastí chromozomu a patrně se s ním i dělí. Integrovaný stav není náhradou homologního genetického materiálu, ale je spíše brán jako doplnění chromozomu. Ve druhém, autonomním stavu, jsou ty epizomy, které se dělí nezávisle a často i rychleji než daná buňka.[11][12][13]

Příkladem epizomu mohou být transpozony a inzerční sekvence nebo faktor F. F faktor určuje, zda je genetický materiál v chromozomu jednoho organismu přenesen do jiného. Transpozony a inzerční sekvence, též nazývány jako mobilní genetické prvky, jsou schopny existovat mimo chromozom, ale po přesunu z jedné buňky do druhé se dokáží na chromozom připojit. Transpozony s sebou nesou další genetický materiál,

který může buňce zajistit např. rezistenci vůči určitým léčivům. Inzerční sekvence též nesou další genetický materiál, ovšem jeho funkcí je pouze připojení ke chromozomální DNA. [1][12][13]

1.2 Mutace

Mutaci můžeme definovat jako kvalitativní nebo kvantitativní změnu v genetické informaci, považuje se za nevratný děj. Přenos genetické informace podléhá náhodným vlivům (mutagenům), např. fyzikálním, chemickým či jiným faktorům. Mezi fyzikální faktory se řadí ionizující nebo UV záření. Jako mutagen mohou působit i viry, které indukují v buňkách mutace. Mutace mohou postihnout zárodečné buňky a přenáší se tak do potomstva (gametické mutace). Pokud jsou mutace tělní a nepřenáší se do potomstva, hovoří se o mutacích somatických. Negativním dopadem mutací může dojít ke změnám v regulačních oblastech transkripce, sekvencích promotorů nebo v signálních sekvencích. [14][15]

Mutace se mohou rozdělit do tří základních skupin podle místa a organizační úrovně jejich vzniku na mutace genové, chromozomové a genomové. Mutace mohou být též slučitelné se životem (vitální) nebo se životem neslučitelné (letální). Ovšem mutace letální nemusí být vždy neslučitelné se životem, jsou to tzv. podmíněné letální mutace, kdy jsou letální pouze v jednom, restriktivním, prostředí, ale slučitelné v jiném, permisivním, prostředí. Tyto mutace jsou nejužitečnější z hlediska genetických studií, protože umožňují studie esenciálních genů. [1][14]

Podle okolností jejich vzniku se dělí mutace na spontánní a indukované. Spontánní mutace vznikají díky chybě při replikaci DNA. Dochází k nim bez zásahu vnějšího prostředí. DNA polymeráza je ovšem velice přesná a má samoopravnou funkci, proto je pravděpodobnost vzniku takových mutací velice malá. Z toho plyne, že naprostá většina mutací vzniká díky působení vnějších faktorů, takové mutace se označují jako indukované. Další definicí mutace může být změna v genotypu organismu oproti normálnímu stavu. Jsou to náhodné změny, s cílenými mutacemi se dá setkat pouze v rámci výzkumu. Z hlediska evoluce jsou mutace velice přínosné, někdy jsou označovány jako tzv. hybná síla evoluce. [14][15][16]

1.2.1 Chromozomové mutace

Chromozomovou mutací se rozumí změna v tvaru nebo struktuře chromozomů. Obecně jsou označovány jako chromozomové aberace. Nejčastější příčinou těchto změn jsou zlomy chromozomu, kdy zlom je přerušení souvislosti DNA řetězce, jehož vinutím chromozom vzniká. Zlomy jsou způsobeny nadměrným působením mutagenů na jedince, či zhoršenou funkcí reparačních mechanismů. Následky aberací jsou závislé na tom, jestli

je zachováno normální množství genetické informace, pokud ano, nazývají se změnami balancovanými. U mutací balancovaných nedochází k fenotypovým projevům. Změna fenotypu může nastat, pokud se zlom vytvoří uprostřed genu nebo mezi kódující sekvencí a jejím promotorem. K fenotypovým projevům může dojít u aberací nebalancovaných, které mají často velice závažné až letální následky. [14][16]

Chromozomové mutace se dělí na inverzi, duplikaci, translokaci a deleci. Inverze nastává, pokud se zlomený chromozomální fragment obrátí a připojí se zpět v obrácené poloze. Duplikace je opakování chromozomálního fragmentu. Při translokaci nastává transfer části chromozomu na jiné místo. Delecí se rozumí ztráta fragmentu chromozomu.[1]

2 KOMPARATIVNÍ GENOMIKA

Komparativní (srovnávací genomika) je vědním oborem zabývajícím se srovnáním genomů různých druhů. Porovnáním kompletních genomových sekvencí jednotlivých organismů může ukázat jejich jednotlivé příbuznosti na genetické úrovni nebo co je naopak od sebe odlišuje. Srovnávací genomika je též využívána při zkoumání evoluce a změn, které nastaly v jejím průběhu, pomáhá při identifikaci genů, které jsou konzervovány nebo se běžně vyskytují mezi druhy, stejně tak i geny dávající organismům jejich jedinečné vlastnosti. [17][18]

Komparativní genomika srovnává vlastnosti jako je velikost genomu nebo počet genů. Porovnávají se i tzv. synteny, což je situace, kdy jsou geny uspořádány v podobných blocích u různých organismů.

2.1 Anotace genomu

Anotace genomu je proces nacházení umístění genů a dalších kódovacích oblastí a určení jejich funkce. V bioinformatice se na základě laboratorně nalezených oblastí snaží předpovídat pozice dalších oblastí pomocí výpočetních metod. Anotaci genomu je možné rozdělit do tři kroků – identifikace částí nekódujících proteiny, genová predikce a připojení biologických informací. Nejjednodušší způsob anotace je vyhledávání homologních genů ve vyhledávači, jako je např. BLAST. Počet informací neustále roste, další poznatky získané anotací objasňují nejasné rozdíly mezi geny, které mají stejnou nebo velmi podobnou anotaci.[17][19]

Anotaci genomu je možno rozdělit na dva druhy, strukturní a funkční. Strukturní anotace shromažďuje informace o identifikaci genomických prvků – ORF (open reading frame – otevřený čtecí rámec) a jejich lokalizace, struktura genu, kódování oblastí aj. Funkční anotace naopak spočívá v přidání biologických informací – biologické a biochemické funkce, exprese a regulace. U prokaryotních organismů se využívají metody založeny na znalostech vlastností promotorů. U eukaryot je použití značně složitější, jelikož strukturně je jejich genom mnohem komplikovanější. [17][19]

3 SYNTENY

Vývoj genomu ovlivňují jak malé bodové mutace v sekvenci DNA, tak i velké události, jako je přeskupení, které přeskládá genetický materiál v buňce. Přeskupení se týká buď jen pár genů, například mutace nebo zlom v důsledku nashromáždění bodových mutací nebo tandemových duplikací. Ale může být i v mnohem větším měřítku, jako jsou dlouhé inverze nebo duplikace celého genomu. Všechny takové události jsou velice důležité v evolučních procesech a plyne z nich, že pokud porovnáme dva a více chromozomů, je velice nepravděpodobné, že pořadí genů bude totožné, dokonce i pro příbuzné druhy. Ovšem pořadí genů není náhodné a při srovnání dvou chromozomů nebo jejich částí, každý od jiného příbuzného druhu, lze nalézt shodné nebo alespoň podobné i celé genové úseky zakonzervované v určitém pořadí. Jak moc jsou si jednotlivé organismy podobné udává míra podobnosti.[20]

Při zjišťování polohy synteny v bakteriálních genomech se využívá různých nástrojů komparativní genomiky, jako jsou například GeneOrder 4.0 a SyMAP. GeneOrder 4.0 je webový nástroj pro analýzu synteny a pořadí genů velkých bakteriálních genomů. Zobrazuje synteny tak, že vykreslí skóre podobnosti proteinů mezi dvěma genomy, též zobrazí anotaci hypotetických proteinů (jejich funkce je prozatím neznámá). Do nástroje se zadávají přístupová čísla jednotlivých genomů v genbank formátu databáze NCBI. Soubory jsou načteny, převedeny na proteinové sekvence a slouží jako vstup do BLAT algoritmu. Díky BLAT algoritmu se mohou analyzovat větší genomy nad 2Mb. Z algoritmu vystupuje hodnota skóre, která určí, jestli jsou sekvence homologní, velmi podobné nebo jen částečně. [21]

SyMAP je systém pro zjištění a zobrazení oblastí synteny pomocí FPC (fingerprint) map. Tento systém počítá synteny bloky a následně je pomocí webové grafiky zobrazuje. Výpočet předpokládaných syntenických sad probíhá pomocí dynamického programování. [22]

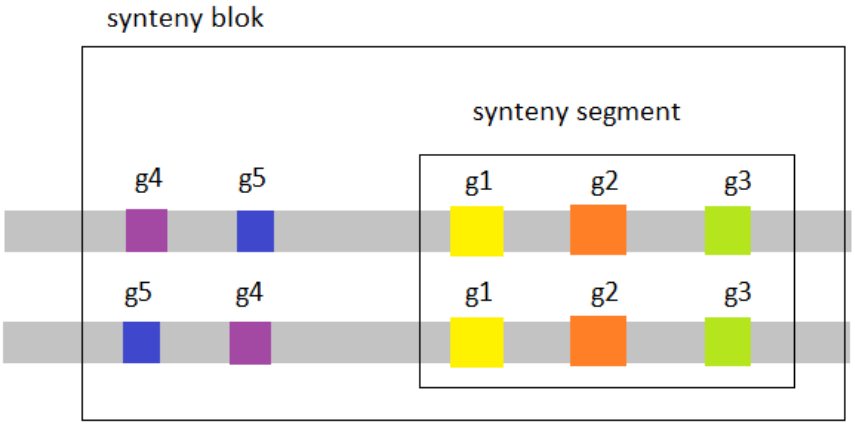
3.1 Synteny bloky

Synteny rozumíme skupinu homologních sekvencí (sekvence odvozeny od stejné původní sekvence, tzn. mají společného předka), které se ve chvíli porovnání genomů vyskytují zároveň. Pokud se porovnávají dva potomci, lze nalézt skupiny po sobě jdoucích sekvencí, ovšem jejich pořadí se může lišit. [20]

Genomové sekvenování a mapování umožnilo srovnání struktur genomů mnoha různých organismů. Zjištěním bylo, že některé organismy mají podobné genové úseky v podobných polohách v genomu. Například mnohé z genů lidí jsou syntenické s těmi

jiných savců – lidoopů, myší atd. Studie synteny může ukázat vývoj genomu v průběhu evoluce. [20]

Pokud jsou místa výskytu sekvence shodné, včetně pořadí, označují se jako synteny segment. Ve chvíli, kdy je větší výskyt synteny segmentů, označuje se tento úsek jako synteny blok. V synteny bloku jsou obsaženy i inverzní a permutované synteny segmenty. V praxi je ovšem vyhledávání synteny mnohem náročnější. Jelikož segment je oblast nalezená ve dvou genomech, kde homologní geny zachovávají stejné pořadí i relativní polohu v genu v obou genomech a savčí genom je složen z 25000 genů o celkové délce $3x10^9$ párů bází. To znamená, že v průměru by segmenty, které obsahují tři geny, přesáhly 300 000 párů bází, což by pro hledání synteny bloků nebylo vhodné. Navíc je velice nepravděpodobné, že při takové délce by sekvence zůstaly naprosto shodné. [20]



Obrázek 3.1 Příklad synteny segment a synteny blok

Proto se využívá takzvaných sekvenčních kotev (anchors). Jako sekvenční kotva je definována konzervovaná neopakující se subsekvence, která se vyskytuje v porovnávaných genomech společně s určitou sekvencí. Počet sekvenčních kotev může být vyšší, než je počet genů. Předpokladem může být subsekvence f_i v genomu F shodná se subsekvencí g_i v genomu G, a zarovnání f_i ke G, dává největší shodu g_i, a naopak. Subsekvence f_i a g_i mohou být použity jako sekvenční kotvy nebo orientační značky (landmarks), které napomáhají v porovnávání dvou genů. V zásadě může být mnohem více sekvenčních kotev, než je genů, což zpřesní rozdělení synteny bloků. Synteny bloky mohou být též definovány pomocí pozičně ukotvených sekvencí (sekvenčních kotev). Je to soubor sekvenčních kotev, které se objevují společně, ovšem ne nutně ve stejném pořadí, v genomu dvou různých organismů. [20]

3.2 Evoluční procesy a synteny bloky

Informace ohledně proběhlých evolučních procesech jsou pro lidstvo velice důležité, proto vzniká snaha o analýzu přestavby různých genomů. Součást jakékoliv analýzy je i řešení permutačních vektorů. Je to snaha o nalezení sledu permutací, jehož výsledkem je shoda v pořadí segmentů v první sekvenci s pořadím, v jakém se nalézají i v sekvenci druhé. Stanovuje se minimální počet permutací a to proto, že se nesnaží nalézt evoluční kroky, ale určit příbuznost jednotlivých druhů – čím nižší počet permutací, tím vyšší příbuznost.[20]

Permutační vektor je množina s *n* prvky, a každý prvek bude představovat úsek DNA (geny v podobě synteny bloku). Na příkladu, kdy *n=5*, se definuje první sekvence jako uspořádaný permutační vektor S_1 = [1 2 3 4 5] a druhá, přeházená sekvence jako S_2 = [2 4 1 5 3]. Cílem teď bude získat shodné pořadí obou vektorů pomocí transformace vektoru S_2. Využívá se metody eliminace bodů zlomu inverzí nebo transpozicí. [20]

Inverze je asi nejběžnější způsob přeskupování, kdy se část vektoru (sekvence) jednoduše převrátí. Tímto způsobem se dá též dosáhnout i uspořádání. Pokud se vezmou sousedící prvky vektoru a jejich diference je rovna 1, pak se berou jako navazující. Ovšem, pokud je diference větší, než je 1, místo se označuje jako bod zlomu. Tato metoda přidává na začátek a konec setřiďovaného vektoru záchytné body, s nimiž není možná manipulace. Počet provedených inverzí by měl být větší nebo roven polovině počtu zlomů a zároveň menší nebo roven počtu bodů zlomu ve vektoru. [20]

Eliminace bodů zlomů pomocí transpozice se zakládá na vyjmutí jednoho nebo více prvků z vektoru a jejich následné přeložení do správné pozice. Použití transpozice je výhodnější, jelikož lze odstranit až tři body zlomu najednou. Nejčastěji se ovšem využívá kombinace obou metod. [20]

3.3 Breakpointová metoda

Breakpointová metoda spadá do skupiny tzv. greedy algoritmů (hladových algoritmů). Využití hladových algoritmů při porovnávání DNA sekvencí bývá většinou mnohem rychlejší než dynamické programování, ale jejich výsledek nemusí být optimální. V každém svém kroku vybírají lokální minimum, s tím, že existuje šance, že takto naleznou minimum globální. Využívá se v případech, kdy se ze zadané množiny vybírá podmnožina s minimální (nebo maximálním) ohodnocením. V případě této práce je to minimální počet kroků nutný k setřídění pozičního vektoru. [23]

Když se použije obecný příklad, na kterém by se hladová strategie dala aplikovat: pokladník uvažuje v každém kroku pouze největší denominaci menší nebo rovnu M. Jelikož cílem bylo minimalizovat počet mincí, které se vrátily zákazníkovi, zdá se to jako velice dobrá strategie. Přeci by se nevrátily nikdy čtyři pětikoruny místo jedné

dvacetikoruny. Tento příklad použil to, co se zdálo jako nejlepší volba a nepovažoval ostatní možnosti. Což je to, co dělá tento algoritmus „hladovým“. Společná vlastnost hladových algoritmů je, že často přináší neoptimální výsledky, za to trvají velmi krátce. Pokladník uvažuje v každém kroku pouze největší denominaci menší nebo rovnu M. [24]

Breakpointová metoda využívaná v této práci minimalizuje body zlomu mezi synteny bloky. Bod zlomu se umisťuje mezi dva bloky (geny), které na sebe nenavazují. Využití počtu zlomů vede k lepšímu algoritmu pro třídění reverzí, protože produkuje řešení bližší optimálnímu. Permutace P_1 až P_n se prodlouží o hodnoty $P_0=0$ a $P_{n+1}=n+1$ na konec, kdy parametr n vyjadřuje maximální hodnotu ve vektoru zvětšenou o jedna. Tyto hodnoty svou polohu v průběhu třídění nikdy nemění. Pokud sousední prvky splňují podmínku, že P_i a P_{i+1} (kdy $0 \leq i \geq n$) jsou po sobě jdoucí čísla, nazývají se přilehlými (adjacency), pokud ne, nazývají se body zlomu (breakpoint). [24]

0 2 1 3 4 5 8 7 6 9

Příklad permutačního vektoru s přidanými pevnými hodnotami na obou koncích

Příklad permutace výše má 5 hodnot přilehlých (2 1 . 3 4 . 4 5 . 7 6) a čtyři body zlomu (0 2, 1 3, 5 8, 6 9). Permutace může mít maximálně $n+1$ bodů zlomu, také nemusí obsahovat žádný, a to v případě, že permutace shodná. Každý bod zlomu odpovídá dvojici prvků permutačního vektoru P_i a P_{i+1}, kde sice tyto dva prvky spolu sousedí v permutačním vektoru P, ale v setříděné matici tomu tak být nemá. Proto jsou nenásledné prvky v třídícím procesu odděleny od sebe a setříděny. Tímto způsobem se dá pozorovat třídění reverzí jako proces odstraňování zlomových bodů. Každá reverze může vést k eliminaci nejvýše dvou breakpointů, což znamená, že:

$$d(P) \geq \frac{b(P)}{2}, \qquad (7.1)$$

kde $b(P)$ je počet bodů zlomu v P. Při setřiďování je vhodné si definovat části permutačního vektoru, které se nachází mezi dvěma po sobě jdoucími body zlomu. Například na příkladu permutace, viz výše, se nachází pět takovýchto úseků (0, 2 1, 3 4 5, 8 7 6, 9). Tyto úseky se nadále mohou hodnotit jako vzestupné (3 4 5) nebo sestupné (2 1, 8 7 6). Jednoprvkové části by se dali klasifikovat buď jako vzestupné nebo i sestupné, ovšem je vhodné je klasifikovat jako sestupnou část, s výjimkou krajních hodnot (0 a $n+1$), které budou vždy rostoucí. [24]

4 GENBANK

GenBank je databáze všech veřejně přístupných anotovaných nukleotidových sekvencí. Databázi spravuje Národní centrum pro biotechnologické informace (NCBI), které je součástí National Institutes of Health (NIH) v USA. K databázi je umožněn přístup přes internet anebo si lze celou databázi bezplatně nainstalovat. GenBank zajišťuje nejaktuálnější obsáhlé informace ohledně nukleotidových sekvencí. Jediný problém databáze je, že do databáze mohou přispívat různí autoři a jejich příspěvky nejsou kontrolovány, proto se nelze plně spoléhat na správnost dat. [25][26]

GenBank formát slouží pro zápis biologických sekvencí z databáze NCBI, jeho přípona je *.gb. Výhodou formátu je, že může obsahovat více záznamů od různých autorů, a též spoustu doplňkových informací, což jiné, takhle využívané formáty, nedokáží. Příkladem je třeba FASTA formát. Samotný GenBank formát se skládá z hlavičky a sekvence nukleotidů. Sekvence je uvozena slovem „ORIGIN“, ukončená znakem „//“.[26][27]

V hlavičce se nachází různé užitečné informace ohledně dané sekvence. Nalezneme zde identifikační číslo sekvence, např. SCU49845, kde první tři znaky označují organismus, čtvrtý a pátý znak přesněji určuje co sekvence kóduje. Dnes již není takové identifikační číslo dostačující pro uvedení všech informací, proto se v současné době přiřazuje identifikační číslo tak, aby bylo jedinečné pro každou sekvenci. Dále v hlavičce nalezneme délku sekvence, což je počet nukleotidových párů bází (dále jen bp). Není zde uveden žádný limit pro velikost vložené sekvence, jediné, co je limitováno, je délka záznamu na minimálně 50 bp. Tento limit byl odsouhlasen pro usnadnění manipulace se sekvenčními daty pomocí softwarových programů. Po délce sekvence je uveden typ molekuly, která byla sekvenována, např. DNA, RNA, transfer RNA, ribosomal RNA. [26][27]

Následuje oddíl GenBank, v hlavičce označen třípísmennou zkratkou názvu oddílu, do které je daný organismus zařazen. Databáze GenBank má 18 oddílů, do kterých můžeme sekvence zařadit – PRI – sekvence primátů (primate sequence), MAM – ostatní savčí sekvence (other mammalian sequences), PLN – sekvence rostlin, hub a řas (plant, fungal and algal sequences), atd. Některé oddíly obsahují sekvence od různých skupin organismů, jiné obsahují data, která byla generována specifickými sekvenčními metodami z mnoha různých organismů. Datum na konci hlavičky určuje den, kdy byla provedena poslední úprava záznamu.[27]

Kromě hlavičky, jsou informace o sekvenci uloženy v metadatech, která jsou součástí formátu GenBank. Stručný popis sekvence nalezneme v části s názvem „DEFINITION“, který zahrnuje jméno organismu, název genu (popř. proteinu), pokud je

sekvence nekódující, tak popis její funkce. Pokud má sekvence kódující úsek (CDS), přidává se poznámka ohledně kompletnosti, například „complete CDS“ (kompletní kódující úsek). Po popisu sekvence následuje přístupové číslo záznamu sekvencí „ACCESSION“. Obvykle je kombinací písmen a číslic, např. jedno písmeno následováno pěti číslicemi. Některá přístupová čísla mohou být delší, v závislosti na délce sekvence. V části „VERSION“ je přístupové číslo doplněno o identifikační číslo verze nukleotidové sekvence, tzn. pokud bude sekvence neupravována od jejího nahrání, bude za přístupovým číslem následovat .1, pokud bude jednou upravena, následovat bude .2, atd. Pokud se změní tohle identifikační číslo, je zároveň sekvenci přiděleno i nové identifikační číslo sekvence „GenInfo Identifier“. Každá proteinová translace má taktéž své vlastní GI číslo, které se při jakékoliv úpravě změní. [27]

V metadatech nalezneme i „KEYWORDS“ neboli klíčová slova, která popisují sekvenci. Tato klíčová slova se ovšem vyskytují pouze ve starších záznamech, v novějších pouze když o to sám autor zažádá nebo záznam obsahuje sekvenci generovanou sekvenční metodou. Dále je zde zdroj „SOURCE“, který obsahuje informace o organismu, ze kterého sekvence pochází. Nachází se tu i informace o autorech, článcích, ve kterých byly dané sekvence zmíněny autory, dále místo, kde byl článek zveřejněn. Důležitou součástí metadat jsou tzv. „FEATURES“, kde jsou informace ohledně genů, genetických produktů, významných úsecích v sekvenci, jako např. oblasti kódující proteiny, mRNA, tRNA, a řadu dalších vlastností. Ve Features se nachází shrnutí informací ohledně sekvence (její délka, zdrojový organismus a taxonomické identifikační číslo), ovšem nejdůležitější jsou informace ohledně CDS (kódující úsek), obsahuje start a stop kodon, překlad do aminokyselin, název vzniklého proteinu. [27]

5 VLASTNÍ ŘEŠENÍ

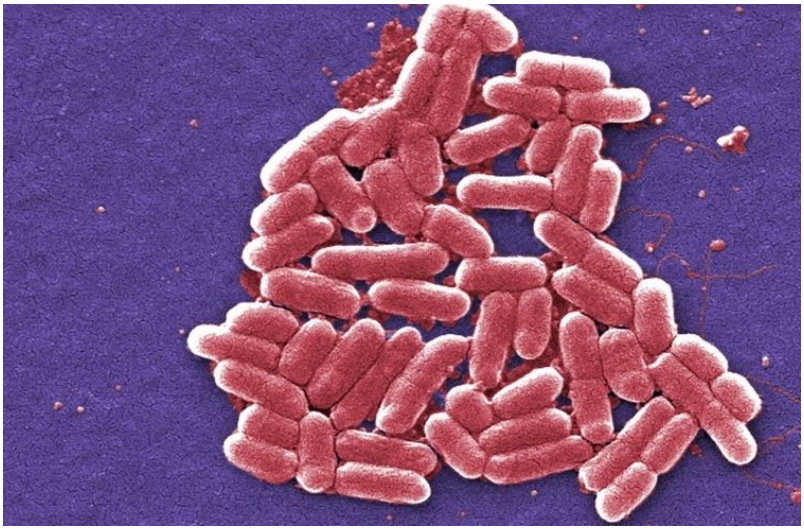
5.1 Dataset bakterií

V následující části práce je popsáno porovnávání anotovaných genomů vybraných bakterií vůči referenční *E. coli*. Soubory bakteriálních genomů jsou stáhnuty z databáze spravované NCBI ve formátu genbank, kde každý soubor má své uvedené databázové číslo, proto je snadné dohledat použité bakterie. Všechny soubory s bakteriálními genomy jsou vybrány na základě označení *RefSeq*, které se nachází v části Keywords a označuje, že sekvence byla anotována pomocí anotace prokaryotického genomu NCBI a měla by zlepšovat konzistenci celého datového souboru. Soubory s označením *RefSeq* by proto měly být ucelenější a kvalitnější. Některé soubory jsou jako referenční sekvence vybrány na základě dlouhodobého využívání a širokého uznání, např. referenční genom Escherichia coli str. K-12 substr. MG1655, který je využíván i v této práci. Všechny funkce budou programovány v prostředí MATLAB. [28]

5.1.1 Escherichia coli

Escherichia coli neboli *E. coli* je použita jako referenční bakterie pro výpočet pozičního profilu genů. *E. coli* je součást fyziologické mikroflóry tlustého střeva a nachází se též v distální části ilea. Vyskytuje se v organismu takřka od narození, nejčastěji alimentární cestou nebo přenosem od jiného jedince, který je *E. coli* už osídlen. *E. coli* není schopna dlouhodobě existovat mimo hostitele.

Kmeny *E. coli* u zdravého jedince nevyvolávají onemocnění, ovšem v případě narušení poměru jednotlivých druhů mikrobů ve střevě mohou způsobit zdravotní komplikace, např. přemnožením *E. coli*. I některé patogenní kmeny mohou způsobovat závažné infekce. Kultivace *E. coli* je nenáročná, nejčastěji se využívá laktózový nebo krevní agar, na kterém roste v šedých koloniích. Enzymatické aktivity se využívá při inkubaci kolonie, jelikož umožňuje přesné zařazení druhu podle jeho biochemických vlastností. *E. coli* patří mezi gramnegativní bakterie. Databázové číslo používané *E. coli* v NCBI databázi je NC_000913. [29]



Obrázek 5.1 *Escherichia coli*, snímek z elektronového mikroskopu (SEM), převzato z [30]

5.1.2 Ostatní použité bakterie

Výběr následujících bakterií byl v důsledku jejich příbuznosti s referenční bakterií *E. coli* nebo dle negativnosti při Gramově barvení. Některé jsou ze stejné čeledi, s těmi další se shodují až ve třídě.

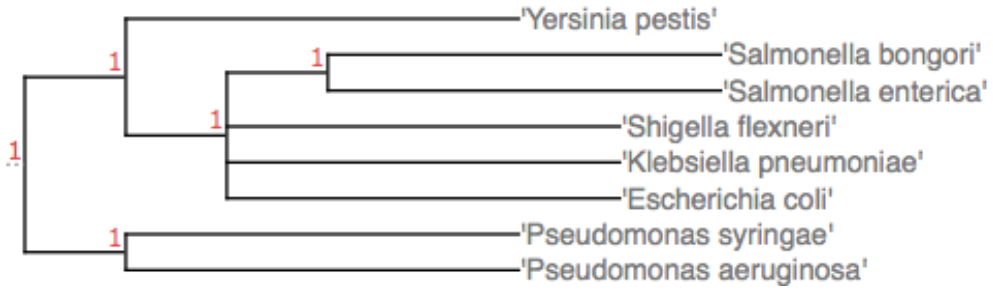
První použitá bakterie je *Klebsiella pneumoniae*, která patří do stejné čeledi *Enterobacteriaceae* jako *E. coli*. Je gramnegativní, nepohyblivá, fakultativně anaerobní, tyčinková bakterie. Vyskytuje se v ústech, kůži a střevech, ovšem pokud se dostane do plic (konkrétně alveolů), může vést až ke krvavému sputu. Ze stejné čeledi jako *E. coli* je též *Shigella flexneri*, která je též gramnegativní, nepohyblivá tyčinka. Způsobuje nemoc shigelózu, což je akutní, vysoce nakažlivé průjmové onemocnění, zdrojem infekce je nejčastěji kontaminovaná potrava. Další použitá bakterie z čeledi *Enterobacteriaceae* je patogenní, gramnegativní, fakultativně anaerobní *Yersinia pestis*. *Y. pestis* je přenositelná jak na zvíře, tak na člověka a způsobuje morovou nemoc. Tato bakterie je důvodem vysoké úmrtnosti během morových epidemií v historii. Infekce touto bakterií se dělí do tří forem: dýmějový, septický a plicní mor. *Salmonella enterica* a *Salmonella bongori* jsou další dvě spadající do jedné čeledi s *E. coli*. Jsou to fakultativně anaerobní, gramnegativní, a hlavně patogenní bakterie. Způsobují gastrointestinální onemocnění nazývané salmonelóza, *S. bongori* způsobuje salmonelózu spíše u plazů, zatímco *S. enterica* je lidským patogenem. [31][32][33]

Bakterie zmíněné výše jsou všechny zařazeny do jedné čeledi, ovšem v této práci jsou využívány ještě další, které s *E. coli* (a ostatními) sdílí jen zařazení do stejné třídy, a tou jsou *Gammaproteobacteria*. *Pseudomonas aeruginosa* a *syringae* jsou gramnegativní, aerobní bakterie. *P. aeruginosa* se nejčastěji vyskytuje v odpadních vodách, na rostlinách a v půdě, je patogenní, může způsobit infekci v lidském těle, např. infekci popálenin nebo oka. *P. syringae* je rostlinný patogen infikující velké množství

druhů a vytváří tzv. bakteriální skvrny. Poslední využívaná bakterie je *Vibrio cholerae*, které též spadá až do stejné třídy. Je to gramnegativní, patogenní bakterie způsobující choleru. [31][32]

Tabulka 5.1 Příbuzné porovnávané bakterie

Jméno bakterie	Délka genomu [bp]	Počet genů	Databázové číslo NCBI
Escherichia coli	4641652	4319	NC_000913
Klebsiella pneumoniae	5520319	5327	NZ_AP014950
Salmonella enterica	4809037	4110	NC_003198
Salmonella bongori	4460105	4267	NC_015761
Shigella flexneri	4607202	4051	NC_004337
Yersinia pestis	4653728	3798	NC_003143
Pseudomonas aeruginosa	6264404	5572	NC_002516
Pseudomonas syringae	6093698	5089	NC_007005
Vibrio cholerae	2961149	2534	NC_002505



Obrázek 5.2 Referenční fylogenetický strom vytvořený pomocí funkce *Common taxonomy tree* na webových stránkách NCBI

5.2 Tvorba pozičního vektoru

Sekvence vybraných bakterií (viz Tabulka 5.1) byly stáhnuty z databáze GenBank, kde byl vybrán celý genom, ne pouze jeho části, které se dají taktéž stáhnout. Příkazem *genbankread* byly nahrány jednotlivé sekvence do prostředí Matlab, ovšem formát genbank obsahuje velké množství v tuto chvíli nepotřebných informací. Potřebné jsou pouze informace o názvech genů. Jelikož referenční *E. coli* obsahuje přes 4000 CDS, byl program nejprve natrénován na části genomu – prvních 50 genů. Pokud by se ponechal celý genom *E. coli*, byl by algoritmus výpočetně náročný, proto byl referenční genom omezen na 1000 genů. Tato hodnota byla vyzkoušena experimentálně na již natrénovaném algoritmu. Nejprve bylo omezení na prvních 500 genech, ale po průchodu algoritmem byla shoda minimální, proto se hranice zvedla na 1000.

Genbank formát neobsahuje všechny názvy hledaných genů, proto pro lepší práci s algoritmem byla prázdná pole nahrazena pouhým znakem ‚X'. Po detailnějším prozkoumání načteného souboru byly nalezeny shodné názvy některých proteinů, označeny jako 'insertion element IS1 protein InsB' a 'insertion element IS1 protein InsA', zkráceně InsB a InsA. Jelikož měly tyto geny stejné označení, stejný produkt a shodné sekvence aminokyselin, byly ze souboru zcela odstraněny. Jejich podobnost by totiž zkreslovala výsledek vyhledávání.

5.2.1 Vyhledávání podobností na základě genů

Hlavním úkolem algoritmu je vyhledat geny, které jsou obsaženy jak v *E. coli,* tak v ostatních bakteriích. Ovšem spolu se shodou je důležitá i informace o pozicích genů, která pomůže určit, jak moc jsou si dané bakterie příbuzné. Proměnná *bakterie* označuje uživatelem vybranou bakterii, která se má spolu s *E. coli* porovnávat. Algoritmus pro vyhledávání shody názvů genů funguje na základě porovnávání, nejprve délek daných slov – to proto, aby se zjednodušila výpočetní náročnost algoritmu, protože pokud název genu z porovnávané bakterie bude mít rozdílnou délku než název genu *E. coli,* cyklus přejde na další iteraci. Z cyklu jsou též odstraněny i ona ‚X', pro jejich nepotřebnost při porovnávání. Poté se porovnává přímo obsah daných proměnných. Výsledkem funkce je třířádkové buňkové pole, kde na prvním řádku jsou názvy genů dle referenční *E. coli* a na řádku druhém pozice jednotlivých genů dle vybrané bakterie a na třetím pozice v referenční bakterii. Zároveň se vytváří i matice jedniček a nul, ve které se nachází jednička na té pozici, na které se našla shoda v názvu genu. Tyto pozice jsou zadány podle *E. coli*.

Tabulka 5.2 Výsledky shody v názvech genů s *E. coli*

Název bakterie	Počet shod
Klebsiella pneumoniae	48
Pseudomonas aeruginosa	226
Pseudomonas syringae	85
Salmonella bongori	40
Salmonella enterica	92
Shigella flexneri	534
Yersinia pestis	362
Vibrio cholerae	102

Z tabulky (Tabulka 5.2) lze vidět, že v některých bakteriích se nalezla mnohem větší shoda než u ostatních. Je to dáno nejen výskytem genů, ale i anotací genomu, některé

soubory bakterií totiž obsahují mnohem více informací v CDS než jiné. Proto přichází na řadu druhá část vyhledávání shody, a to pomocí lokálního zarovnání aminokyselin.

5.2.2 Vyhledávání podobností na základě translace

Vytvořená pomocná matice jedniček a nul v této části je velmi nápomocná pro zlepšení výpočetní náročnosti, nalezené shody budou vynechány z algoritmu lokálního zarovnání. Smithův – Watermanův algoritmus provádí lokální zarovnání, což je hledání nejpodobnějších úseků různých délek mezi dvěma sekvencemi. Oblasti, které jsou od těchto úseků vzdáleny, nejsou při zarovnání brány v potaz. Lokální zarovnání povoluje mezery, je to algoritmus dynamického programování, který sestavili pánové Temple F. Smith a Michael S. Waterman v roce 1981. Základem lokálního zarovnání je Needleman – Wunschův algoritmus pro globální zarovnání, ovšem lokální zarovnání pokládá negativní hodnoty rovny nule. [34] Jako skórovací matice, která penalizuje vložení a prodloužení mezery, byla použita v Matlabu pro proteiny BLOSUM80, jelikož se zpracovávají velmi podobné sekvence.

Když je shoda nalezena v názvech genů, dále se nepoužívají. Vybírají se pouze sekvence aminokyselin obsažené v CDS, u kterých se provádí lokální zarovnání pomocí funkce *swalign*. Výstupem této funkce je zarovnání sekvencí a jejich skóre. Minimální hranice skóre byla nastavena tak, že se pro každou sekvenci aminokyselin v *E. coli* spočítala maximální možná hodnota skóre (kdy by byly obě sekvence totožné) a ta se pak snížila její 80% hodnotu. Maximální skóre bylo získáno ze skórovací matice BLOSUM80, kde se na diagonále nachází skóre pro totožné znaky, jednotlivé sekvence byly příkazem *aa2int* převedeny z aminokyselin na čísla a vypočítána nejvyšší možná hodnota skóre pro každou sekvenci. Snížení na 80% hodnotu bylo nastaveno experimentálně, s touto hranicí se dá manipulovat. V matici, kde je uloženo skóre zarovnání se najde maximální hodnota určující podobnost sekvence aminokyselin *E.coli* k sekvenci aminokyselin, která byla vzata z porovnávané bakterie. Díky skórovací matici a určení minimální hranice skóre se zvyšuje pravděpodobnost, že sekvence si budou opravdu podobné a geny tedy správně určeny jako shodné. Tento algoritmus je ovšem velice náročný, jak výpočetně, tak časově. Omezení na prvních 1000 genů vyhledávání sice urychlí, ale i tak jeden cyklus vyhledávání shod trval v průměru dvě hodiny.

Nejlepší výsledky porovnání byly pro bakterii *Shigella flexneri*, která obsahovala dostatečné množství informací o názvech genů i shodu v zarovnání translace. Shoda se našla v 534 názvech genů a 201 lokálních zarovnání translací, což je 735 shod z celkového 1000. Na těchto výsledcích lze pozorovat, že *Shigella* a *Escherichia* jsou velice podobné bakterie.

Jelikož programovací prostředí Matlab umí rozpoznat v sekvenci aminokyselin pouze základních 20 označení pro aminokyseliny, ostatní znaky, ač též jsou rozšířením označení pro aminokyseliny, Matlab nedokáže vyhodnotit a vypíše chybovou hlášku. Proto bylo potřebné tyto znaky ze sekvence odstranit.

Tabulka 5.3 Porovnání výsledků vyhledávání shod pro jednotlivé bakterie

Název bakterie	Shody v názvech	Shody v translaci	Celkem shod
Klebsiella pneumoniae	48	480	528
P. aeruginosa	226	0	226
P.syringae	85	2	87
Salmonella bongori	40	562	602
Salmonella enterica	92	552	644
Shigella flexneri	534	201	735
Yersinia pestis	362	31	393
Vibrio cholerae	102	16	118

5.3 Třídění pozičního vektoru

Pro setřídění pozičního vektoru jednotlivých bakterií byla použita breakpointová metoda, kdy se z CDS vybrala položka *indices*, která přestavuje počáteční a konečnou pozici sekvence genu. Vybrána byla pouze počáteční pozice, pro zjednodušení se neuvažovalo, jestli se gen nachází na hlavním nebo komplementárním vlákně. Před použitím breakpointové metody se musely data upravit, protože se v každé bakterii nalezly shody v různých genech a pro další zpracování je nezbytné, aby všechny poziční vektory měly stejnou délku. Proto se všechny výsledky shod porovnaly, a pokud se gen nacházel ve všech, ponechal se. Genů, které se nalezly ve všech bakteriích, bylo celkem 60. Tato hodnota platí jen pro tento data set bakterií s použitými parametry. Počet společných genů by byl vyšší, kdyby se například snížila minimální hodnota skóre, ale za cenu toho, že by se mohly i méně podobné sekvence označit jako shodné.

Jelikož se v pozičním vektoru vyskytovalo velké množství osamocených hodnot, které nebyly součástí ani sestupné ani vzestupné sekvence, bylo nutné před samotným použitím breakpointové metody tyto hodnoty setřídit. Nejprve byly do vektoru přidány ukotvené hodnoty, 0 na začátek, délka vektoru+1 na konec. Byla vytvořena funkce, která kontrolovala, zda se před nebo za každou hodnotou nachází číslo, které se liší o jedna. Pokud ne, tyto hodnoty byly vybrány a z pozičního vektoru odstraněny.

Příklad pozičního vektoru:	0 . 4 3 . 5 . 1 2 . 9 . 7 6 . 8 . 10
Odstranění osamocených hodnot:	0 . 4 3 . 1 2 . 7 6 . 10
Osamocené hodnoty:	5, 9, 8

Osamocené hodnoty byly setříděny vzestupně a postupně zpět zařazeny do pozičního vektoru tak, že se nalezla hodnota o jedna menší, než je hodnota vybrané osamocené hodnoty. Při třídění osamocených hodnot bylo třeba dodržet návaznost v pozičním vektoru, například hodnota 5 by byla seřazena mezi 0 a 4, aby byla dodržena klesající posloupnost.

Zařazení osamocených hodnot:	0 . 5 4 3 . 1 2 . 8 7 6 . 9 10

Po zařazení osamocených hodnot do jednotlivých posloupností, byl zjištěn jejich charakter, tedy zda jsou vzestupné či sestupné. Vzestupné části byly v pomocném vektoru nahrazeny nulami a sestupné části jedničkami. Důležité bylo nalézt i první nesetříděnou hodnotu (první bod zlomu), kdy v tomto příkladu se nachází hned mezi 0 a 5. Z vybraných sestupných posloupností byla vybrána ta s nejmenší hodnotou na jejím konci, tedy 5 4 3. Od místa výskytu prvního bodu zlomu až po tuto minimální hodnotu byla provedena inverze. Znovu se zjistí pozice sestupných částí a místo prvního bodu zlomu a cyklus pokračuje do chvíle, dokud není celý vektor setříděn.

První krok:	0 . 3 4 5 . 1 2 . 8 7 6 . 9 10
Druhý krok:	0 . 6 7 8 . 2 1 . 5 4 3 . 9 10
Třetí krok:	0 1 2 . 8 7 6 5 4 3 . 9 10
Čtvrtý krok:	0 1 2 3 4 5 6 7 8 9 10

Ojediněle ale může nastat případ, že ve vektor po třetím setřídění dostane do stejného tvaru, v jakém byl již po prvním, a program se zacyklí. Proto bylo nutné si pamatovat předchozí kroky setřiďování. Pokud se stane, že se jednotlivé vektory budou shodovat a hrozilo by zacyklení programu, je situace řešena inverzí pouze sestupné sekvence s nejnižší hodnotou. Poté program běží dále.

Výstupem breakpointové metody je počet osamocených hodnot a počet kroků nutný k setřídění vektoru. Pokud tyto dvě hodnoty sečteme, dostaneme finální počet kroků nutný k seřazení vektoru breakpointovou metodou. Tyto kroky znázorňují evoluční vzdálenost mezi *E. coli* a vybranou bakterií. Čím menší bude počet kroků potřebný k seřazení, tím jsou bakterie příbuznější.

Tabulka 5.4 Počty kroků pro setřídění pozičního vektoru pro jednotlivé bakterie

Název bakterie	Počet osamocených hodnot	Počet kroků metody
Shigella flexneri	0	0
Klebsiella pneumoniae	3	1
Salmonela bongori	0	0
Salmonella enterica	0	0
Pseudomonas syringae	15	16
Pseudomonas aeruginosa	13	16
Yersinia pestis	6	6
Vibrio cholerae	11	28

5.4 Diskuze výsledků

Výstupem breakpointové metody je počet kroků nutný k setřídění pozičního vektoru. Jako další parametr pro vyhodnocení vztahů mezi bakteriemi se může využít *p*-distance neboli proporcionální vzdálenost. Její klasicky používaný výpočet je podíl počtu mutací v sekvenci a její délky, ovšem pro vyhodnocení se vzorec upraví na podíl počtu kroků breakpointové metody n_p a délky sekvence n (7. 1). Čím menší je hodnota *p*-distance, tím více jsou si sekvence podobné.

$$p = \frac{n_p}{n} \qquad (7.\ 1)$$

Tabulka 5.5 Seznam hodnot *p*-distancí pro jednotlivé bakterie

Název bakterie	Počet kroků	Délka sekvence	*p*-distance
Shigella flexneri	0	60	0
Klebsiella pneumoniae	4	60	0.0667
Salmonela bongori	0	60	0
Salmonella enterica	0	60	0
Pseudomonas syringae	31	60	0.5167
P. aeruginosa	29	60	0.4833
Yersinia pestis	12	60	0.2
Vibrio cholerae	39	60	0.65

Z tabulky (Tabulka 5.5) lze jasně vidět, že nejpodobnější bakterie k *E. coli* jsou *Shigella flexneri* a obě *Salmonelly*. Nulové hodnoty v počtu kroků jsou dány tím, že finální počet genů vstupujících do breakpointové metody byl pouze 60, bakterie jsou sice velice příbuzné, ale ne shodné. Bakterie by měly mít mnohem více společných genů, ale jelikož do algoritmu vstupovalo z referenční bakterie pouze prvních 1000 genů, je možné, že další shodné geny se nachází mimo tento rozsah.

Vyhledávání shod pomocí názvů bylo pro některé bakterie velice úspěšné, nalezená shoda u *Shigelly flexneri* byla 534 genů z necelého 1000 genů (byla odmazána prázdná a opakující se pole). U dalších bakterií toto vyhledávání nebylo tak úspěšné, nejméně pro *Salmonellu bongori*, kde se shodovalo jen 40 názvů. Tyto výsledky jsou nejspíše dány nedostatečnou anotací genomů, jelikož ve spoustě genomů nebyly názvy genů uvedeny. Vyhledávání na základě lokálního zarovnání sekvence aminokyselin pro bakterie *Klebsiella pneumoniae, Salmonella bongori a enterica*, bylo nejúspěšnější a shodovala se cca půlka genů. Žádná shoda nebyla nalezena u *Pseudomonas aeruginosa*, nejspíše kvůli nastavené hranici minimálního skóre zarovnání, protože při experimentálním zjišťování této hodnoty, kdy byla hranice nejprve nastavena jako 70% procent z maximální možné hodnoty skóre, se nalezlo shod 14.

Z výše zobrazené tabulky (Tabulka 5.5) lze odhadnout pomocí proporcionální vzdálenosti podobnost a evoluční vzdálenost bakterií. Jasně se dá říci, že nejpodobnější bakterie k referenční *Eschrichii coli* jsou *Shigella flexneri* a *Salmonella enterica* a *bongori*. Za nimi následuje *Klebsiella pneumoniae*, která by měla dle referenčního fylogenetického stromu (Obrázek 5.2) podobnější než zmíněné *Salmonelly*. Další by byla podle *p*-distance *Yersinia pestis*, pak *Pseudomonas aeruginosa* a *syringae*, a nakonec *Vibrio cholerae*. Tato posloupnost bakterií seřazená podle evoluční vzdálenosti odpovídá taxonomickému rozdělení fylogenetického stromu, proto se dá říci, že výstup algoritmu je správný.

6 ZÁVĚR

V této bakalářské práci bylo hlavním cílem vytvoření algoritmu pro výpočet pozičního profilu genů. Objasnění základních pojmů jako je bakteriální genom, plazmidy a mutace se zabývá první kapitola. V kapitole druhé jsou zmíněny pojmy jako anotace genomu a komparativní genomika. Třetí kapitola se zabývá pojmy synteny, synteny bloky a obsahuje teoretický popis breakpointové metody, která je v práci využívána. V úvodu do praktické části jsou zmíněny použité bakterie a jejich vlastnosti. Bakterie byly vybrány dle společné vlastnosti – gramnegativní barvení, též jsou všechny zařazeny do stejné třídy. Jako referenční bakterie byla použita *Escherichia coli*, jejíž genom je jeden z nejvíce prozkoumaných a hojně se využívá i ve výzkumu. Některé porovnávané bakterie způsobují různá onemocnění (*Shigella, Salmonella, Yersinia, Vibrio*) nebo i bakterie vyskytující se v odpadních vodách nebo parazitující na rostlinách.

Další část práce se zabývá vytvořením algoritmu pro výpočet pozičního profilu genů anotovaných genomů vůči referenčnímu genomu. Zvoleny byly dvě úrovně vyhodnocování podobností na základě informací obsažených v genbank záznamu sekvencí vybraných bakterií. Výstup vyhledávání je dále zpracováván breakpointovou metodou a výsledky jsou interpretovány v poslední kapitole. Jako nejpříbuznější bakterie byly určeny *Shigella flexneri, Salmonella enterica a bongori*, což je dle taxonomického zařazení správné.

Literatura

[1] SNUSTAD, D. Peter a Michael J. SIMMONS, RELICHOVÁ, Jiřina, ed. *Genetika*. Přeložil Anna MATALOVÁ. Brno: Masarykova univerzita, 2009. ISBN 978-80-210-4852-2.

[2] *Genetika – Biologie* [online]. [cit. 2018-01-27]. Dostupné z: http://www.genetika-biologie.cz/prokaryota

[3] GEORGE P. RÉDEI. *Encyclopedia of genetics, genomics, proteomics, and informatics*. 3rd ed. New York: Springer, 2008. ISBN 9781402067532.

[4] KAGUNI, Jon M. DnaA: Controlling the Initiation of Bacterial DNA Replication and More. *Annual Review of Microbiology* [online]. 2006, **60**(1), 351-371 [cit. 2018-01-30]. DOI: 10.1146/annurev.micro.60.080805.142111. ISSN 0066-4227. Dostupné z: http://www.annualreviews.org/doi/10.1146/annurev.micro.60.080805.142111

[5] BRYANT, John A. a Stephen J. AVES. Initiation of DNA replication: functional and evolutionary aspects. *Annals of Botany* [online]. 2011, **107**(7), 1119-1126 [cit. 2018-08-06]. DOI: 10.1093/aob/mcr075. ISSN 1095-8290. Dostupné z: https://academic.oup.com/aob/article-lookup/doi/10.1093/aob/mcr075

[6] MOTT, Melissa L. a James M. BERGER. DNA replication initiation: mechanisms and regulation in bacteria. *Nature Reviews Microbiology*[online]. 2007, **5**(5), 343-354 [cit. 2018-08-06]. DOI: 10.1038/nrmicro1640. ISSN 1740-1526. Dostupné z: http://www.nature.com/articles/nrmicro1640

[7] YAKOVCHUK, P. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research*[online]. 2006, **34**(2), 564-574 [cit. 2018-08-06]. DOI: 10.1093/nar/gkj454. ISSN 0305-1048. Dostupné z: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkj454

[8] EDITED BY DONALD R. HELINSKI, STANLEY N. COHEN, DON B. CLEWELL, DAVID A. JACKSON a Alexander HOLLAENDER. *Plasmids in Bacteria*. Boston, MA: Springer US, 1985. ISBN 9781461324478.

[9] SNYDER, Larry. a Larry. SNYDER. *Molecular genetics of bacteria*. 4th ed. Washington, DC: ASM Press, c2013. ISBN 1555816274.

[10] *Bacterial DNA – the role of plasmids* [online]. [cit. 2018-01-03]. Dostupné z: https://www.sciencelearn.org.nz/resources/1900-bacterial-dna-the-role-of-plasmids

[11] EDITED BY NANCY L. CRAIG ... [ET AL.]. *Mobile DNA II*. Washington, D.C: ASM Press, 2002. ISBN 9781555817954.

[12] HRUBAN, Vojtěch a Ivan MAJZLÍK. *Obecná genetika*. Praha: Česká zemědělská univerzita, 2000. ISBN 978-80-213-0600-4.

[13] CAMPBELL, Allan M. *Episomes* [online]. Elsevier, 1963, 1963, s. 101-145 [cit. 2018-08-06]. Advances in Genetics. DOI: 10.1016/S0065-2660(08)60286-2. ISBN 9780120176113. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0065266008602862

[14] KOLEKTIV, Oldřich Nečas a. *Obecná biologie pro lékařské fakulty*. 3., přeprac. vyd., V nakl. H. Jinočany: H, 2000. ISBN 80-86022-46-3.

[15] *Genotoxicita a karcinogeneze* [online]. [cit. 2018-01-30]. Dostupné z: https://is.muni.cz/do/rect/el/estud/prif/ps13/genotox/web/pages/02_mutace.html

[16] *Genetika – Biologie* [online]. [cit. 2018-01-30]. Dostupné z: http://www.genetika-biologie.cz/mutace

[17] TOUCHMAN, Jeffrey. Comparative Genomics. *Nature Education Knowledge* [online]. 2010 [cit. 2018-08-06]. Dostupné z: https://www.nature.com/scitable/knowledge/library/comparative-genomics-13239404

[18] CVRČKOVÁ, Fatima. *Úvod do praktické bioinformatiky*. Praha: Academia, 2006. ISBN 80-200-1360-1

[19] *National Human Genome Research Institute* [online]. [cit. 2018-01-31]. Dostupné z: https://www.genome.gov/11509542/

[20] RICHARD C. DEONIER, MICHAEL S. WATERMAN a Simon TAVARÉ. *Computational Genome Analysis An Introduction*. New York, NY: Springer Science+Business Media, 2005. ISBN 0387288074.

[21] MAHADEVAN, Padmanabhan a Donald SETO. Rapid pair-wise synteny analysis of large bacterial genomes using web-based GeneOrder4.0. *BMC Research Notes* [online]. 2010, **3**(1), 41- [cit. 2018-08-06]. DOI: 10.1186/1756-0500-3-41. ISSN 1756-0500. Dostupné z: http://bmcresnotes.biomedcentral.com/articles/10.1186/1756-0500-3-41

[22] SODERLUND, C. SyMAP: A system for discovering and viewing syntenic regions of FPC maps. *Genome Research* [online]. 2006, **16**(9), 1159-1168 [cit. 2018-08-06]. DOI: 10.1101/gr.5396706. ISSN 1088-9051. Dostupné z: http://www.genome.org/cgi/doi/10.1101/gr.5396706

[23] ZHANG, Zheng, Scott SCHWARTZ, Lukas WAGNER a Webb MILLER. A Greedy Algorithm for Aligning DNA Sequences. *Journal of Computational Biology* [online]. 2000, **7**(1-2), 203-214 [cit. 2018-08-07]. DOI: 10.1089/10665270050081478. ISSN 1066-5277. Dostupné z: http://www.liebertpub.com/doi/10.1089/10665270050081478

[24] JONES, Neil C a Pavel PEVZNER. *An introduction to bioinformatics algorithms*. Cambridge, MA: MIT Press, c2004. ISBN 0-262-10106-8.

[25] *GenBank Overview* [online]. [cit. 2018-01-03]. Dostupné z: https://www.ncbi.nlm.nih.gov/genbank/

[26] BENSON, D. A., I. KARSCH-MIZRACHI, D. J. LIPMAN, J. OSTELL a D. L. WHEELER. GenBank. *Nucleic Acids Research* [online]. 2007, **35**(Database), D21-D25 [cit. 2018-01-03]. DOI: 10.1093/nar/gkl986. ISSN 0305-1048. Dostupné z: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkl986

[27] *Sample GenBank Record* [online]. [cit. 2018-01-03]. Dostupné z: https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

[28] *Prokaryotic RefSeq Genomes* [online]. [cit. 2018-05-14]. Dostupné z: https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/

[29] HORÁČEK, Jiří. *Základy lékařské mikrobiologie*. Praha: Karolinum, 2000. ISBN 80-246-0006-4.

[30] HANEY CARR, Janice. *Bacteria in photos* [online]. [cit. 2018-08-07]. Dostupné z: http://www.bacteriainphotos.com/Escherichia%20coli%20electron%20microscopy.html

[31] RYAN, Kenneth J, C. George RAY a John C SHERRIS. *Sherris medical microbiology: an introduction to infectious diseases*. 4th ed. New York: McGraw-Hill, c2004. ISBN 0-8385852-9-9.

[32] BENEŠ, Jiří. *Infekční lékařství*. Praha: Galén, c2009. ISBN 978-80-7262-644-1.

[33] GERARD J. TORTORA, BERDELL R. FUNKE a CHRISTINE L. CASE. *Microbiology: an introduction*. 9th ed. New Delhi (India): Pearson, 2009. ISBN 8131722325.

[34] SMITH, T. F. a M. S. WATERMAN. *Identification of Common Molecular Subsequences* [online]. 1981, , 195-197 [cit. 2018-05-21]. Dostupné z: https://dornsife.usc.edu/assets/sites/516/docs/papers/msw_papers/msw-042.pdf

SEZNAM PŘÍLOH

Příloha 1: Obsah přiloženého CD

1. Bakalářská práce, formát .pdf
2. Složka „Algoritmy a bakterie"
    - soubory bakterií ve formátu .gb
    - hlavní skript: Skript_Martinkova.m
    - pomocné funkce ve formátu .m